CHAPTER 4

STRUCTURING A DATA LAKE FOR THE MANAGEMENT OF SCIENTIFIC INFORMATION IN BRAZIL

Washington Luís Ribeiro de Carvalho Segundo
Brazilian Institute of Information in Science and Technology (IBICT), Brazil.
ORCID: *https://orcid.org/0000-0003-3635-9384*

Fábio Lorensi do Canto
Central Library, Federal University of Santa Catarina, Brazil.
ORCID: *https://orcid.org/0000-0002-8338-1931*

Patrícia da Silva Neubert
Department Information Science, Federal University of Santa Catarina, Brazil.
ORCID: *https://orcid.org/0000-0002-8909-1898*

Adilson Luiz Pinto
Department Information Science, Pós-Design, Federal University of Santa Catarina, Brazil.
ORCID: *https://orcid.org/0000-0002-4142-2061*
Email: *adilson.pinto@ufsc.br*

Carlos Luis González-Valiente
Publications Department, Pro-Metrics, Tallinn, Estonia.
ORCID: *https://orcid.org/0000-0002-1836-5257*

ABSTRACT

The initial steps involved in the establishment of a data lake (Laguna) were delineated. This data lake was fed with structured data from the data ecosystem of the Brazilian Current Research Information System (BrCris). The data lake was developed to manage scientific information and aggregate this content into an accessible system. A substantial amount of data was collected and processed across five phases: (1) collection; (2) selection and separation; (3) transformation and connection; (4) organization, classification, and indexing; and (5) retrieval and visualization. The study utilized a range of data extraction methodologies on disparate platforms, employing SQL or API to facilitate the process. A set of scientific journals was identified through a process of stratification, with the highest percentage belonging to the A1 category. The initial integration of OpenAlex and DOAJ data was conducted, marking a significant milestone in the development of the platform. The author data were disambiguated and cross-checked by DOI to identify citing and cited authors. A comprehensive set of relevant data was obtained to facilitate the formulation of robust inferences, including the standardized number of journals by stratification, the integration between disparate databases such as OpenAlex and DOAJ, the ontological system employed to address the disassociation of authors, and the representation of the cited author before journals and future authorities.

KEYWORDS: data gap, data interoperability, scientific information management, academic databases

HOW TO CITE: Luís Ribeiro de Carvalho Segundo, W., Lorensi do Canto, F., Neubert, P. da S., Luiz Pinto, A., & González-Valiente, C. L. (2025). Structuring a data lake for the management of scientific information in Brazil. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 122-143). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.112.

1 INTRODUCTION

This research is supported by two projects: Laguna de Datos and the Brazilian Current Research Information System (BrCris; Pinto et al., 2022). These projects are overseen by the research group Brazilian Scientific Research Ecosystem Laboratory (LAEPECBR, in Portuguese). The objective of this initiative is to establish a data lake structure within Brazil, with the aim of supporting open data systems and ecosystems within the Brazilian Institute for Research in Science and Technology (IBICT, in Portuguese; Dias et al., 2022; Segundo et al., 2022). A significant challenge confronting Brazilian science and technology institutions pertains to the heterogeneity of the data recovered, characterized by a lack of cohesive structures. The construction of the IBICT data lake structure aims to address this challenge (Segundo & Sena, 2023). Concurrently, it aspires to function as a dependable repository for novel research data derived from the BrCris project. BrCris is an ecosystem that encompasses a comprehensive array of information and scientific findings. It employs sophisticated algorithms to derive indicators and metrics from recommendation systems, facilitating the identification of four distinct categories of specialists and specialties (Figure 1).

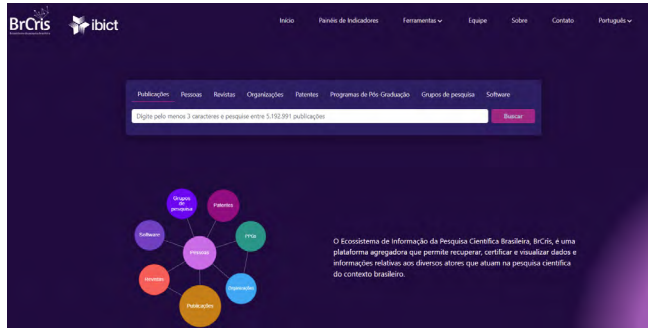

Figure 1. BrCris' webpage. **Source**. https://brcris.ibict.br

Laguna is defined as a system or repository of data stored in its natural format without processing (Ravat & Zhao, 2019). This system constitutes a unified repository of data derived from processed, statistical, and social systems. The objective of this system

is to transform, use, replicate, analyze, learn, and visualize the data in a manner that is accessible to all individuals involved in the systems, including system builders, operators, and users (Nargesian et al., 2019; Oliveira & Martins, 2022). The concept of context encompasses structured, semi-structured, and unstructured data, as well as image, audio, and video data. This type of data is referred to as "binary data" (Silberschatz et al., 2011). The data lake structure is predicated on six levels: (1) a management layer, (2) data access, (3) data collection tools, (4) various data repositories, (5) databases, and (6) a dashboard system so that the community has access to the contents (John & Misra, 2017; Valles-Coral et al., 2023). A data lake is defined as an extensive collection of datasets that can be stored in different systems (Giebler et al., 2019; Gontijo et al., 2021; Netto & Pinto, 2022). It is important to note that these data may be presented in a variety of formats and subject to change over time. The system generates autonomous operating systems, more appropriate reports, and predictive analyses for institutional needs. The data lake is a system that serves to structure a set of data according to its format specification, breakage of contents, content reformat, and data format. It also establishes dataset instances and connections, generates systems to qualify data and their contents, schedules, views, and accesses various data content (Coimbra & Dias, 2021; Segundo & Sena, 2023; Sousa & Shintaku, 2022). This particular data lake is employed by a diverse array of institutions, including companies, governments, and scientific-technological agencies.

The primary objective of its implementation is to ensure that the data are presented and stored in a scalable structure, with execution systems organized in clusters. These systems are required to process and store a substantial volume of data concurrently, in addition to executing these processes with open-source software and independently of file size. These systems are designed to collect potential data, identify user needs, monitor user behavior, detect fraud and data risks, manage marketing systems, analyze competition, and customize data demand. The objective of this initiative is to establish a data repository, designated as Laguna, which will serve as a foundation for the BrCris ecosystem. The specific objectives are as follows: (1) to incorporate statistical data, data aggregation APIs, and visualize certain scientific inferences, (2) to categorize scientific journals according to the Brazilian class identification model (Qualis/Capes), (3) to

identify open-access data crossover, (4) to perform author disambiguation for a more accurate representation of scientific data, and (5) to cross-reference the data by DOI to identify the degree of citations.

2 DESIGN AND METHODOLOGY

This research employed sophisticated computational methodologies for the management, structuring, and examination of information. This was done to obtain searchable, accessible, interoperable, and reusable datasets. The dataset has been meticulously collected, selected, transformed, and linked for the purpose of data processing. Subsequently, the data were methodically organized and indexed, and then retrieved and rendered visually on the BrCris platform. As a mining processing technique, six phases of data crossing were used: (1) understanding the scientific environment, (2) understanding of the data to be worked on, (3) preparation of these data, (4) mathematical modeling of these data, (5) evaluation of possible applicable models, and (6) generation of a data production system.

2.1 Data treatment model

The data lake is defined as a collection of data that has been meticulously gathered from repositories and databases of recognized pertinence within the domains of science, technology, and innovation. The study focused on the direct engagement with the localization, accessibility, interoperability, and reuse of datasets, aligning with the FAIR Principles (Wilkinson et al., 2016). The following sources were consulted: OpenAlex, Wikidata, CrossRef, OpenCitations, OpenAIRE Research Graph, ISSN Portal, Latindex, DOAJ, Google Scholar Metrics, Plataformas Lattes and Sucupira, Oasisbr, and BDTD. The datasets obtained from these sources were then subjected to various analytical procedures, including advanced artificial intelligence techniques, real-time analytics, machine learning algorithms, dashboards, and visualizations. As the main results, we obtained:

1. *Collection*: We employed public APIs and search and extraction tools available in repositories and databases that fully or partially complied with the FAIR Principles. In addition, tools were developed for the purpose of extracting data from complex sources (do Carmo & da Silva Lemos, 2022). The utilized protocols encompassed the transmission and reception of messages through Representational State Transfer (REST) interfaces, with calls facilitated by HyperText Transfer Protocol (HTTP). This was done to obtain responses from documents in JSON format. Furthermore, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) protocol is employed, with HTTP calls and responses in the form of eXtensible Markup Language (XML) in accordance with various standards. In accordance with the tenets of the OAI—Dublin Core, we have obtained responses in more generic models, such as the Resource Description Framework (RDF), which exhibits a high degree of expressiveness.

2. *Selection and separation*: The subsequent filtration and categorization processes were then executed. Ancillary information for the collection was eliminated. The collected files were dismembered into the different types of entities described in their content, which, in this study, are referred to as "payloads."

3. *Transformation and connection*: Adaptations and validations were generated, as well as the establishment of relationships with records from other sources. A record obtained from source A exhibited a shared attribute with a record from source B, thereby enabling the establishment of a link between the two with a certain degree of reliability. The integration of the other record attributes resulted in the formation of a comprehensive, unified record. The process of duplication was eradicated.

4. *Organization, classification, and indexing*: The data that underwent classification served as the foundation for constructing search interfaces, web services, and visualization panels. The retrieval process involved the use of search facets in unstructured textual fields, encompassing full text, through actions such as tokenization and stemming.

5. *Retrieval and visualization*: The indicators were developed to facilitate the visualization of these metrics on the dashboards. The utilization of visualization tools resulted in the generation of collaboration networks, geospatial data, time series, and dynamic tabulation schemes. The semantic model employed in this study adhered to international standards. These systems were found to be compatible with those employed in other countries to achieve advanced levels of interoperability.

3 RESULTS

The results of the study are derived from the information sources outlined in the methodological framework. The SQL extractions used to identify journal priorities by the Brazilian graduate system were as follows:

3.1 Sucupira Platform to identify the journals and the potential of their indicators

```
SELECT *
FROM (
SELECT sources.abbreviated_title, sources.alternate_titles,
        sources.apc_prices, sources.apc_usd, sources.cited_by_
        count, sources.country_code, sources.counts_by_year,
        sources.created_date, sources.display_name, sources.
        homepage_url, sources.host_organization, sources.host_
        organization_lineage, sources.host_organization_name,
        sources.id, sources.ids, sources.is_in_doaj, sources.is_oa,
        sources.issn_l, sources.publisher, sources.publisher_id,
        sources.societies, sources.summary_stats, sources.type,
        sources.updated_date, sources.works_api_url, sources.
        works_count, sources.x_concepts, qualis.issn, qualis.titu-
        lo, qualis.area_avaliacao, qualis.estrato
FROM laguna.sources, laguna.qualis
WHERE sources.type = 'journal'
AND array_contains(sources.issn, qualis.issn)
) LIMIT 5
```

3.2 *Google Metrics to identify scientific journals' h5 mean and median*

```
root
 |---title: string (nullable = true)
 |---issn: string (nullable = true)
 |---doi_example: string (nullable = true)
 |---title_gsm_2022: string (nullable = true)
 |---h5_2022: string (nullable = true)
 |---med_h5_2022: string (nullable = true)
 |---url_2022: string (nullable = true)
 |---title_gsm_2021: string (nullable = true)
 |---h5_2021: string (nullable = true)
 |---med_h5_2021: string (nullable = true)
 |---url_2021: string (nullable = true)
 |---title_gsm_2020: string (nullable = true)
 |---h5_2020: string (nullable = true)
 |---med_h5_2020: string (nullable = true)
 |---url_2020: string (nullable = true)
 |---title_gsm_2019: string (nullable = true)
 |---h5_2019: string (nullable = true)
 |---med_h5_2019: string (nullable = true)
 |---url_2019: string (nullable = true)
 |---title_gsm_2018: string (nullable = true)
 |---h5_2018: integer (nullable = true)
 |---med_h5_2018: integer (nullable = true)
 |---url_2018: string (nullable = true)
 |---title_gsm_2017: string (nullable = true)
 |---h5_2017: string (nullable = true)
 |---med_h5_2017: string (nullable = true)
 |---url_2017: string (nullable = true)
 |---title_gsm_2016: string (nullable = true)
 |---h5_2016: string (nullable = true)
 |---med_h5_2016: string (nullable = true)
 |---url_2016: string (nullable = true)
 |---title_gsm_2015: string (nullable = true)
 |---h5_2015: string (nullable = true)
 |---med_h5_2015: string (nullable = true)
 |---url_2015: string (nullable = true)
 |---title_gsm_2014: string (nullable = true)
 |---h5_2014: string (nullable = true)
```

```
|---med_h5_2014: string (nullable = true)
|---url_2014: string (nullable = true)
|---title_gsm_2013: string (nullable = true)
|---h5_2013: string (nullable = true)
|---med_h5_2013: string (nullable = true)
|---url_2013: string (nullable = true)
```

3.3 *OpenAlex to identify annual, 2-year, i10, and APC payments*

```
root
|---display_name: string (nullable = true)
|---issn: string (nullable = false)
|---issn_l: string (nullable = true)
|---2yr_mean_citedness: double (nullable = true)
|---h_index: long (nullable = true)
|---i10_index: long (nullable = true)
|---currency: string (nullable = false)
|---price: string (nullable = false)
|---apc_usd: long (nullable = true)
|---country_code: string (nullable = true)
```

With regard to the other databases, data recovery is achieved through extraction in CSV or JSON. As the primary source of information, we have OpenAlex, which is developing a journal recommendation system within BrCris, the final expression of the scientific index process. Given the substantial volume of data, we have implemented a data transformation process with Laguna, whereby we extract the raw data from their various sources and subsequently integrate them into BrCris. This approach was adopted to ascertain the most pertinent journals in the field. For instance, in the context of searching for a subject within the domain of journals (bibliometrics), there is often a multitude of variations in titles, which can pose a significant challenge. The Laguna mechanism facilitates this process, particularly through the incorporation of indicator options that are integrated into the system. Suppose that the search came up with 10 journal titles, such as *Em Questão, Informação & Sociedade: Estudos, Perspectivas em Ciência da Informação, Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Transinformação, Movimiento, Revista de*

Pesquisa Cuidado é Fundamental Online, *AtoZ: Novas Práticas em Informaçãoo e Conhecimento*, *Ciência e Saude Coletiva*, and *Revista de Saúde Pública*, and we need to know which one has the most publications on the subject. Subsequently, these metrics can be utilized in decision-making processes, with considerations given to the h5 mean or median, APC, and response time from the submission of the article to the receipt of an acceptance or rejection. The indices, which have been completed by Laguna and subsequently migrated to BrCris, can assist in determining a recommendation system. In the context of a system for journals, this technology can be utilized by editors to enhance the indexing of their periodicals. With regard to technological systems, such as patents, there is potential for enhanced accessibility for scientists.

The crux of the issue with the Laguna system is the absence of data transformation metrics, as the data are derived directly from the original sources. However, certain components of these sources are utilized, leading to the observed limitations. For instance, in the context of OpenAlex, the system has already developed a subset of the relevant metrics. The proposed solution involves the incorporation of recommendation systems. The study utilizes data from Google Scholar Metrics and the quality system of the Sucupira Platform. The ultimate objective of this study is to create indicators with the data recovered from information sources of magazines, theses, patents, and editorial content:

- *Scientific production*: We have some indices that can be managed, such as: (1) response time from submission to publication, (2) average H-5, (3) median H-5, (4) publication rate, (5) journal editorial committee—national or international, (6) inbreeding for Brazilian journals, and (7) languages of publication.

- *Theses*: We have the indices of (1) genealogy up to 11 levels, (2) regional orientation, (3) parents scientifically, (4) subject matter experts in guidance, and (5) thematic specialists in position shares.

- *Patents*: We can look at the system by (1) citations received, (2) quotes made, (3) concession and renewal, (4) patent family, (5) triadic, and (6) classification.

- **Editorial systems:** We have: (1) average H-5, (2) median H-5, (3) degree of endogamy, (4) production of doctors, (5) production of teachers, and (6) APC billing.

The visualization of this data is represented by priority indices within the BrCris dashboard, which is the output of all these indicators.

3.4 Scientific journal

First, the records of scientific journals were cross-referenced using data extracted from OpenAlex and part of the contents of the Sucupira Platform, which is related to the Brazilian journal evaluation system Qualis/Capes. The enriched set of journal data was integrated into the BrCris platform, rendering it accessible for consultation and data visualization via a file. The file designates the journals as A1, A2, A3, A4, B1, B2, B3, B4, and C, thereby assigning them importance commensurate with that of the aforementioned journals. Figure 2 illustrates the distribution of journal classification strata within the Qualis system.

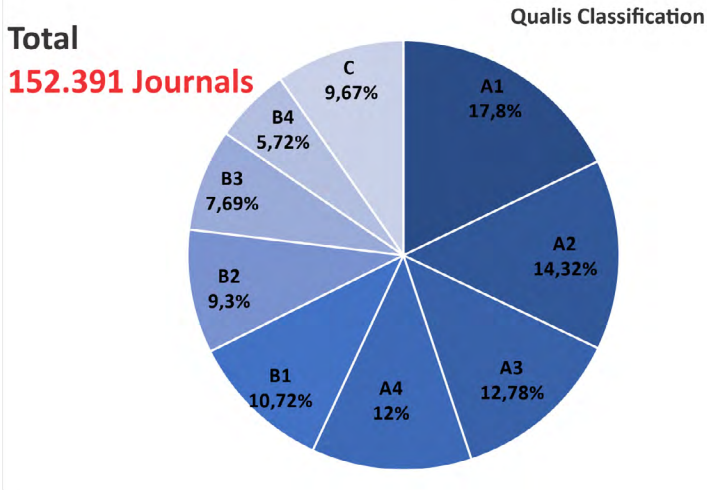

Figure 2. Dashboard of the journals in BrCris. **Source**. *https://brcris.ibict.br/vivo/revistas.*

3.5 *Open-access data crossing*

A cross-referencing process was conducted between data from OpenAlex and DOAJ and data from journals evaluated in the Qualis/Capes system. The variables that were identified during the course of this study included the percentage of open-access journals and the cost of APC by stratum and evaluation area. This dataset represents the initial phase of a cost foresight project that aims to develop a model of transformative agreements for Brazilian scientific production. A set of journals was selected for this study, and five impact indicators were previously calculated using data from open platforms. The indicators employed in this study include the 2yr_mean_citedness (average citations per article in 2 years), the h-index, and the i10 index, which were extracted from OpenAlex. Additionally, the h5 index and the h5 median were retrieved from Google Scholar Metrics. Subsequently, data analysis from open-access journals was carried out by cross-referencing data between the Sucupira, DOAJ, and OpenAlex platforms. The percentage of open access and the price of APC rates of the journals evaluated in Qualis/Capes (2017–2020) were identified, and the final data were quantified by stratum and evaluation area (Witt & Silva, 2022). Table 1 presents the mean APC price of journals by Qualis evaluation stratum. It has been observed that the APC price is elevated in journals of stratum A, particularly in stratum A1, which predominantly encompasses journals of heightened prestige and scientific rigor.

Table 1. APC price of Qualis journals by evaluation stratum.

Qualis evaluation stratum	Average price of APC (USD)
A1	3,403.66
A2	1,970.80
A3	1,834.80
A4	1,339.99
B1	849.38
B2	942.09
B3	1,264.95
B4	714.58
C	1,358.25

Table 2. Presents the mean APC price of journals by Qualis evaluation area.

Evaluation areas	Average price of APC (USD)
Anthropology/Archeology	2,675.68
Astronomy/Physics	2,669.72
Biological Sciences II	2,564.35
Materials	2,552.01
Computation Science	2,551.37
Biological Sciences III	2,547.07
Biological Sciences I	2,546.87
Engineering IV	2,518.90
Chemical	2,503.10
Political Science and International Relations	2,493.76
Engineering III	2,490.44
Mathematics/Probability and Statistics	2,490.40
Pharmacy	2,478.94
Engineering II	2,477.49
Medicine I	2,465.79
Medicine II	2,458.12
Medicine III	2,451.72
Architecture, Urbanism, and Design	2,444.74
Economy	2,422.27
Engineering I	2,392.18
Collective Health	2,370.94
Biotechnology	2,356.84
Interdisciplinary	2,355.33
Psychology	2,344.39
Public and Business Administration	2,305.44
Physical Education	2,300.72
Geosciences	2,287.76
Nutrition	2,260.39

Evaluation areas	Average price of APC (USD)
Biodiversity	2,249.48
Food Science	2,236.70
Sociology	2,221.95
Art	2,217.97
Veterinary Medicine	2,217.24
Dentistry	2,206.19
Agricultural Sciences I	2,197.00
Geography	2,194.22
Law	2,187.16
Nursing	2,184.24
Communication and Information	2,166.03
Environmental Sciences	2,162.37
Linguistics and Literature	2,118.91
Teaching	2,041.29
History	2,018.79
Zootechnism/Fishing Resources	2,000.49
Urban and Region Planning/Demography	1,871.05
Philosophy	1,808.08
Education	1,780.46
Sciences of Religion and Theology	1,719.75
Social Service	1,232.25

The observation revealed that the average price of APCs was notably higher in journals within the domains of Exact and Natural Sciences, Engineering, and Technology. Conversely, in the Human Sciences, Social Sciences, and Arts, the APC cost of journals is lower. Furthermore, a cross-referencing process was conducted between the h5 index of Google Scholar Metrics, OpenAIRE, and DOAJ. The initial dataset, pertaining to journals indexed in OpenAIRE, encompasses 15,366 titles, and the distribution of their h5-index values is illustrated in Figure 3.

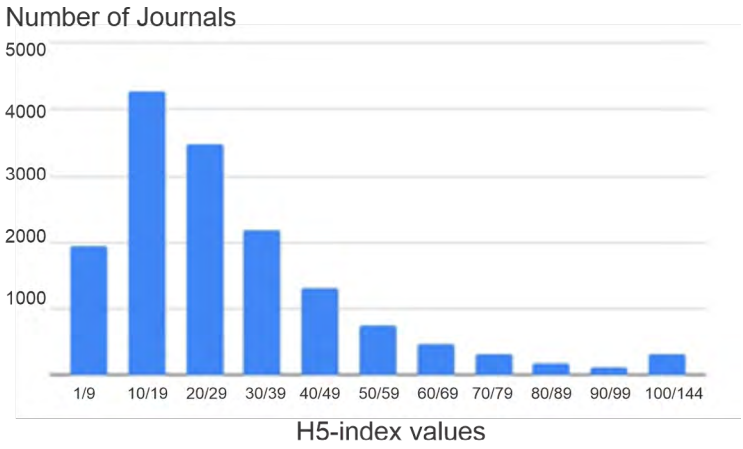

Figure 3. H5 index of journals indexed in OpenAIRE.

The second set included 9,722 open-access journals indexed in DOAJ (Vilas Boas et al., 2023), with ranges of h5-index values described in Figure 4. These data can be used to prepare studies on open access and transformative agreements in Brazil, as well as possible activities to be carried out in the project's subsequent phases.

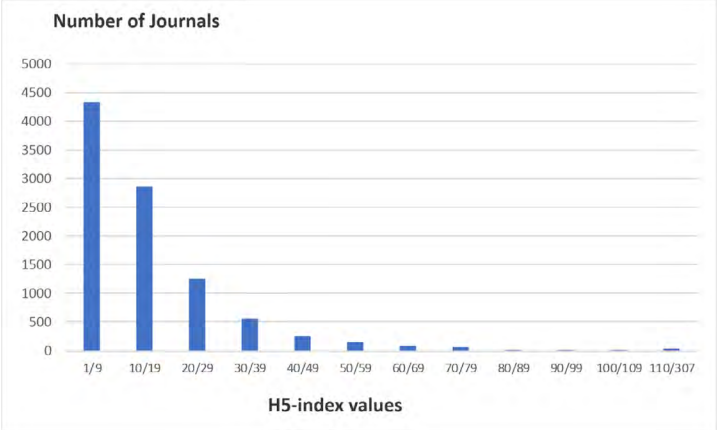

Figure 4. H5 index of journals indexed in DOAJ.

3.6 *Author disambiguation*

A significant challenge in the management of extensive collections of academic and scientific records pertains to the process of name disambiguation. To address this challenge, OpenAlex data are being integrated with the Lattes Platform, a Brazilian database that contains researchers' CVs (Mascarenhas et al., 2021). In the initial data cross-reference, approximately 100,000 authors with ORCID records in both sources were identified and extracted. This process facilitates the validation of author records and the expansion of the collaboration networks presented in BrCris, as illustrated in Figure 5.

Figure 5. Semantic visualization of authors in BrCris.

3.7 Data crossing by DOI

Finally, OpenAlex data were extracted from nearly 5 million papers registered with DOIs in Lattes syllabi. The objective of this study was to analyze the existing data in order to identify a system of citations and cited references. The analysis entailed the identification of several key elements, including cross-citations of editors, the impact of journal citations, and the academic development of certain institutions within high-impact journals. The release of AWS credits from the call of the National Council for Scientific and Technological Development (CNPq, in Portuguese) initiated the training of project team members to configure Laguna's cloud infrastructure. The Laguna data were subsequently uploaded to the cloud server, and testing commenced using AWS processing and analysis tools. Among the tests carried out, the SageMaker notebook demonstrated a notable performance. This is a tool for developing machine learning models. The efficacy of the tool was ascertained through a citation analysis model that focused on scientific journals (Figure 6) and the subsequent generation of a network (Figure 7).

```
In [15]: entity_filter = 'concepts.id:'
         year_filter = 'publication_year:'
         year = '2022'

         # For demonstration purpose we are not going to do pagination
         ai_works = requests.get(openalex_api+'works?filter='+entity_filter+information_system_id+','+year_filter+year).json()
         df_ai_works = pd.DataFrame(ai_works['results'])
         df_ai_works
```

9	https://openalex.org/W4213072884	https://doi.org/10.1016/j.techsoc.2022.101937	Go cashless! Determinants of continuance inten...	Go cashless! Determinants of continuance inten...	2022	2022-02-01	{'openalex': 'https://openalex.org/W4213072884...	en
10	https://openalex.org/W4283378156	https://doi.org/10.1016/j.techfore.2022.121562	Analysis of the adoption of emergent technolog...	Analysis of the adoption of emergent technolog...	2022	2022-05-01	{'openalex': 'https://openalex.org/W4283378156...	en
11	https://openalex.org/W4210681325	https://doi.org/10.1007/s12553-022-00639-w	Semantic interoperability in health records st...	Semantic interoperability in health records st...	2022	2022-01-26	{'openalex': 'https://openalex.org/W4210681325...	en
12	https://openalex.org/W4212841316	https://doi.org/10.1016/j.autcon.2022.104171	Design and implementation of a smart infrastru...	Design and implementation of a smart infrastru...	2022	2022-04-01	{'openalex': 'https://openalex.org/W4212841316...	en
13	https://openalex.org/W4220725585	https://doi.org/10.3390/su14063508	Antecedents and Impacts of Enterprise Resource...	Antecedents and Impacts of Enterprise Resource...	2022	2022-03-16	{'openalex': 'https://openalex.org/W4220725585...	en
14	https://openalex.org/W4205549242	https://doi.org/10.1016/j.jclepro.2022.130557	Green innovations, supply chain integration	Green innovations, supply chain integration	2022	2022-03-01	{'openalex': 'https://openalex.org	en

Figure 6. Figure 6. Test of the scientific journal citation analysis model.

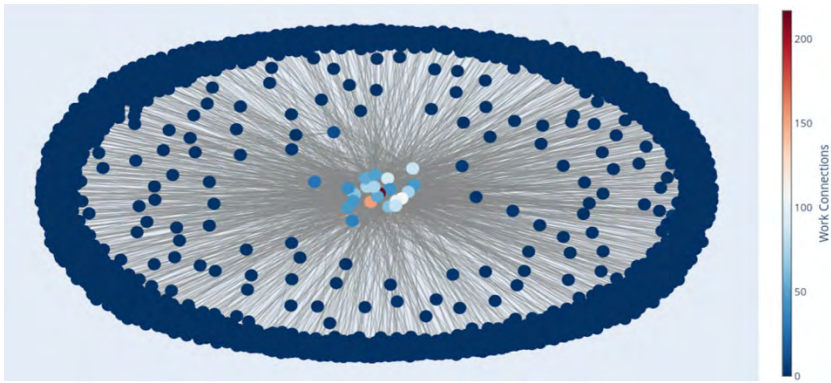

Figure 7. Network generated in testing the scientific journal citation analysis model.

4 CONCLUSIONS

In this study, we constructed the attributes constructor and proceeded to formulate a data lake model. The objectives that were previously outlined have been met, and the data serve as the foundation for the BrCris. The data lake is utilized for the processing of substantial datasets, encompassing both rationing and metrics. The development of certain APIs facilitated the extraction process. Consequently, the development of recommendation systems and the visualization of information in graph models and analytical graphs have become a reality. These initial findings may serve as a foundation for the development of novel research opportunities in the field of Brazilian ecosystems. The subsequent dataset is anticipated to be more extensive and dynamic, incorporating artificial intelligence and machine learning to facilitate the automated processing and indexing of data for aggregation within the BrCris framework. In contemplating the imminent future, it is anticipated that this technology will facilitate the generation and organization of data, thereby enabling the expeditious acquisition of information essential for effective decision-making. This initiative will be expanded to

encompass a broader range of disciplines, including patent data, technical production, and other scientific information types. The advantage of BrCris is that, regrettably, it is incapable of processing a significant amount of data, a capability that is possessed by Laguna. This phenomenon can be attributed to the fact that the systems in question—both hardware and software—were designed with this specific purpose in mind. The system is equipped with an AWS account, which facilitates the hosting of data across all levels. The Laguna serves to unify data in Brazilian science and technology, given that suppliers and services employ divergent methods. The primary function of the data lake is to develop the harmony of these data, a flexible model, and real-time machine learning. This approach ensures the accessibility and currency of the data, facilitating compatibility across a range of platforms, systems, programs, and tools.

Funding

This study was supported by the Brazilian Institute of Information in Science and Technology.

Conflict of interest

The author(s) declare that there is no conflict of interest.

Contribution statement

Conceptualization: Washington Luís Ribeiro de Carvalho Segundo.
Methodology and Data Curation: Fábio Lorensi do Canto.
Formal Analysis: Patrícia da Silva Neubert.
Writing – Original Draft: Washington Luís Ribeiro de Carvalho Segundo, Adilson Luiz Pinto.
Writing – Review & Editing: Carlos Luis González-Valiente, Fábio Lorensi do Canto, Adilson Luiz Pinto.
Supervision: Washington Luís Ribeiro de Carvalho Segundo, Patrícia da Silva Neubert.

REFERENCES

Coimbra, F. S., & Dias, T. M. R. (2021). Use of open data to analyze the publication of articles in scientific events. *Iberoamerican Journal of Science Measurement and Communication*, *1*(3), 1–13. *https://doi.org/10.47909/ijsmc.123*

Dias, T. M. R., Mena-Chalco, J. P., Segundo, W. L. R. C., Pinto, A. L., & Moreira, T. H. J. (2022). BrCris: Plataforma Para Integração, Análises E Visualização De Dados Técnicos-Científicos. *Informação & Informação*, *27*, 622–638. *https://doi.org/10.5433/1981-8920.2022v27n3p622*

do Carmo, D., & da Silva Lemos, D. L. (2022). Padrões de qualidade para dados e metadados endereçados a aplicações em ciência de dados. In *Advanced notes in information science* (vol. 2, pp. 161–170). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.116*

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Leveraging the data lake: Current state and challenges. In C. Ordonez, I. Y. Song, G. Anderst-Kotsis, A. Tjoa, I. Khalil (Eds.), *Big Data analytics and knowledge discovery. DaWaK 2019. Lecture Notes in Computer Science* (p. 11708). Springer. *https://doi.org/10.1007/978-3-030-27520-4_13*

Gontijo, M. C. A., Hamanaka, R. Y., & de Araujo, R. F. (2021). Research data management: A bibliometric and altmetric study based on dimensions. *Iberoamerican Journal of Science Measurement and Communication*, *1*(3), 1–19. *https://doi.org/10.47909/ijsmc.120*

John, T., & Misra, P. (2017). *Data lake for enterprises: Lambda architecture for building enterprise data systems*. Packt Publishing.

Mascarenhas, H., Rodrigues Dias, T. M., & Dias, P. (2021). Academic mobility of doctoral students in Brazil: An analysis based on Lattes Platform. *Iberoamerican Journal of Science Measurement and Communication*, 1(3), 1–15. *https://doi.org/10.47909/ijsmc.53*

Nargesian, F., Zhu, E., Miller, R., & Pu, Q. (2019). Lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment*, (2), 1986–1989. *https://doi.org/10.14778/3352063.3352116*

Netto, M. C. S., & Pinto, A. L. (2022). O silêncio dos dados diz muito, basta prestar atenção: breves experimentos sobre análise exploratória visual. In T. M. R. Dias (Ed.), *Informação, Dados e Tecnologia. Advanced Notes in Information Science* (vol. 2, pp. 15–23). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.118*

Oliveira, L. F. R., & Martins, D. L. (2022). Coleta de dados para agregação de repositórios digitais: Entidades vinculadas à Secretaria Especial de Cultura do Brasil. In *Advanced Notes in Information Science* (vol. 2, pp. 171–181). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.106*

Pinto, A. L., Segundo, W. L. R. C., Dias, T. M. R., Silva, V. S., Gomes, J., & Quoniam, L. M. (2022). Brazil Developing Current Research Information Systems (BrCRIS) as data sources for studies of research. *Iberoamerican Journal of Science Measurement and Communication*, *2*(1), 1–12. *https://doi.org/10.47909/ijsmc.135*

Ravat, F., & Zhao, Y. (2019). Data lakes: Trends and perspectives. In S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. Tjoa, & I. Khalil (Eds.), *Database and Expert Systems Applications. DEXA 2019. Lecture Notes in Computer Science* (pp. 11706, 304–313). Springer. *https://doi.org/10.1007/978-3-030-27615-7_23*

Segundo, W. L. R. C., & Sena, P. (2023, April 5–6). Laguna—FAIR research data infrastructure and open science support observatory [Conference session]. Expert Finder Systems, Coral Gables Miami.

Segundo, W., Dias, T. M., Moreira, T., Pinto, A. L., Silva, V., Gomes, J., Quoniam, L., Matas, L., Dias, A., & Schneider, J. (2022). Uma estratégia para coleta, integração e tratamento de dados científicos no contexto do BrCris. In T. M. R. Dias (Ed.), *Informação, Dados e Tecnologia. Advanced Notes in Information Science* (vol. 2, pp. 215–222). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.117*

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2011). Relational database design. In *Database design concepts* (6th ed.). McGraw-Hill.

Sousa, R. P. M., & Shintaku, M. (2022). Política de privacidade de dados: observações relevantes para sua implementação. In T. M. R. Dias (Ed.), *Informação, Dados e Tecnologia. Advanced Notes in Information Science* (vol. 2, pp. 82–91). ColNes Publishing. *https://doi.org/10.47909/anis.978-9916-9760-3-6.112*

Valles-Coral, M., Injante, R., Hernández-Torres, E., Pinedo, L., Navarro-Cabrera, J. R., Salazar-Ramírez, L., Cárdenas-García, Á., & Huancaruna, E. (2023). Aggregation of institutional repositories for the analysis of the scientific performance of Peruvian universities. *Iberoamerican Journal of Science Measurement and Communication*, 3. *https://doi.org/10.47909/ijsmc.63*

Vilas Boas, R. F., Campos, F. F., Andrade, D. A. F., & Canto, F. L. (2023). Revistas científicas registradas no DOAJ: análise a partir do Índice H5. *BiblioCanto*, *9*(2), 100–115. *https://doi.org/10.21680/2447-7842.2023v9n21D33680*

Wilkinson, M. D., Dumontier, M., Jsbrand Jan Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. *https://doi.org/10.1038/sdata.2016.18*

Witt, A. S., & Silva, F. C. C. da. (2022). Analysis of citizen science in Brazil: A study of the projects registered in the Civis platform. *Iberoamerican Journal of Science Measurement and Communication*, *2*(3). *https://doi.org/10.47909/ijsmc.162*